הטכניון – מכון טכנולוגי לישראל
מעבדה במערכות הפעלה 046210
תרגיל בית מס' 2

תאריך הגשה: 31.7.2024, עד 23:55

Introduction

In the previous assignment you have implemented a simple Message Passing Interface (MPI) between processes. In this assignment you will learn about the scheduling algorithm and how to wake up processes.

Your mission in this assignment is to add a polling mechanism to the MPI system. You are required to add the system call **mpi_poll**, which would be a blocking system call. That is, the system call should not return immediately, but instead it should make the calling process wait the wanted messages are received or a timeout expires.

Working Environment

You will be working on the same REDHAT 8.0 Linux virtual machine, as in the previous assignments.

Detailed Description

This assignment builds on the previous assignment. For this work you should use the MPI system you implemented in the previous assignment and implement a polling algorithm for it.

When a process issues the system call **mpi_poll** the kernel should check if any of the given PIDs has sent a message. If such messages exist, they will be reported. If not, the process that issued **mpi_poll** is put to sleep until any of the requested messages are received. The wait time is limited by a timeout to avoid being put to sleep indefinitely.

The system call should implement this interface:

int mpi_poll(struct mpi_poll_entry * poll_pids, int npids, int timeout)

a. Description:
   Check all entries of the **poll_pids** array (indices 0 to **npids**-1). For each entry **poll_pids[i]**, if there's an incoming message from **poll_pids[i].pid**, set **poll_pids[i].incoming** to 1. Otherwise, set it to zero. If none of the given PIDs had an incoming message, go to sleep until a message arrives (and report it in the correct **poll_pids** entry) or until **timeout** expires.

   A message is considered available by the polling mechanism until **mpi_receive** is called, which receives the message and deletes it from the incoming messages queue. If multiple messages are available after waking up, report all of them.

   The definition of **struct mpi_poll_entry** is as follows. Add it to **mpi_api.h**:

```
struct mpi_poll_entry {
      pid_t pid;
      char incoming;
}
```

   The new system call should use the number 246.

b. Return value
    i. on failure (no messages reported): -1
    ii. on success: The number of entries in **poll_pids** with incoming=1. **poll_pids** should be updated accordingly.

c. On failure **errno** should contain one of following values:
    i. “ETIMEDOUT” (Polling timed out): No message arrived before timeout expired.
    ii. “EFAULT” (Bad address): Error copying **poll_pids** from or to user space.
    iii. "ENOMEM" (No memory): Error allocating memory.
    iv. “EINVAL” (Invalid argument) **npids**<1 or **timeout**<0.
    v. “EPERM” (Operation not permitted): The current process isn’t registered for MPI.

Useful Information

- You can assume that the system is with a single CPU.
- An introduction to task scheduling is given in the Background Information section below
- More on system calls and task scheduling can be found in the “Understanding The Linux Kernel” book.
- Use **printk** for debugging.
- You are not allowed to use **syscall** functions to implement code wrappers, or to write the code wrappers for your system calls using the macro **_syscall1**. You should write the code wrappers according to the example of the code wrapper given above.
- Use Bootlin (see link in Moodle) to easily find where some function/variable is defined in the kernel

Tips for the Solution

- To put a process to sleep, you need to force the kernel not to schedule it. One possible way to do this is to change the status of the process to TASK_INTERRUPTIBLE and set a timeout after which it will be woken. Setting a timeout can be done using the **schedule_timeout**() function (inside the timer.c file). Note that this function calls the **schedule**() function therefore you should not call it from inside the **schedule**() function itself. If you want to initiate a timeout from inside the **schedule**() function, you will need to manually set up a timer (see how it is done in the schedule_timeout() function).
- Take the time to think where to put your changes to avoid breaking existing functionality.

Testing Your Custom Kernel

You should test your new kernel thoroughly (all functionality and error messages that you can simulate). Note that your code will be tested using an automatic tester. This means that you should pay attention to the exact syntax of the wrapper functions, their names and the header file that defines them. You can use whatever file naming you like for the source/header files that implement the system calls themselves, but they should compile and link using the kernel make file. To do so add your source file in the following line inside the “Makefile” file located in the “kernel” folder:

```
obj-y    = sched.o dma.o … <your_file_name>.o
```

Submission Procedure

1. You should submit through the moodle website (one submission per pair).
2. You should submit one zip file containing:
   a. All files you added or modified in your custom kernel (including relevant files from the previous exercise). The files should be arranged in folders that preserve their relative path to the root path of the kernel source, i.e:
   ```
   zipfile -+
            |
            +- submitters.txt
            |
            +- mpi_api.h
            |
            +- kernel/ -+
            |           |
            |           +-...
            |
            +- include/ -+
            |            |
            |            +-...
           ...
   ```
   b. The wrapper functions file "mpi_api.h" that includes, in addition to the system calls from the previous exercise, the wrapper for **mpi_poll** and the definition of **struct mpi_poll_entry**.
   c. A file named "submitters.txt" which lists the name **email** and ID of the participating students. The following format should be used:

      ploni almoni ploni@t2.technion.ac.il 123456789
      john smith john@gmail.com 123456789

      Note: that you are required to include your email.

Emphasis Regarding Grade

- Your grade for this assignment makes 35% of final grade.
- Your submissions will be checked using an automatic checker, pay attention to the submission procedure.
- You are allowed (and encouraged) to consult with your fellow students but you are not allowed to copy their code. You can share tests.
- Your code should be adequately documented and easy to read.
- Incorrect submission format and late submissions will be penalized.
- Obviously, you should free all dynamically allocated memory.

Background Information

This assignment requires basic understanding of the task scheduling in the Linux kernel. Below is some information to get you started. More information can be found in the recommended links in the lab Moodle and the OS course.

Linux is a multitasking operating system. A multitasking operating system achieves the illusion of concurrent execution of multiple processes, even on systems with single CPU. This is done by switching from one process to another very quickly. Linux uses **Preemptive Multitasking.** This means that the kernel decides when a process is to cease running and a new process is to begin running. Tasks can also intentionally **block** or **sleep** until some event occurs (keyboard press, passage of time etc.). This enables the kernel to better utilize the resources of the system and give the user a responsive feeling.

Task States

The state field of the process descriptor describes what is currently happening to the process. The process can be in one of the following states:

TASK_RUNNING: The process is either executing on a CPU or waiting to be executed.

TASK_INTERRUPTIBLE: The process is suspended (sleeping) until some condition becomes true. Raising a hardware interrupt, releasing a system resource the process is waiting for, or delivering a signal are examples of conditions that might wake up the process (put its state back to TASK_RUNNING).

TASK_UNINTERRUPTIBLE: Like TASK_INTERRUPTIBLE, except that delivering a signal to the sleeping process leaves its state unchanged. This process state is seldom used.

TASK_STOPPED: Process execution has been stopped; the process enters this state after receiving a SIGSTOP, SIGTSTP, SIGTTIN, or SIGTTOU signal.

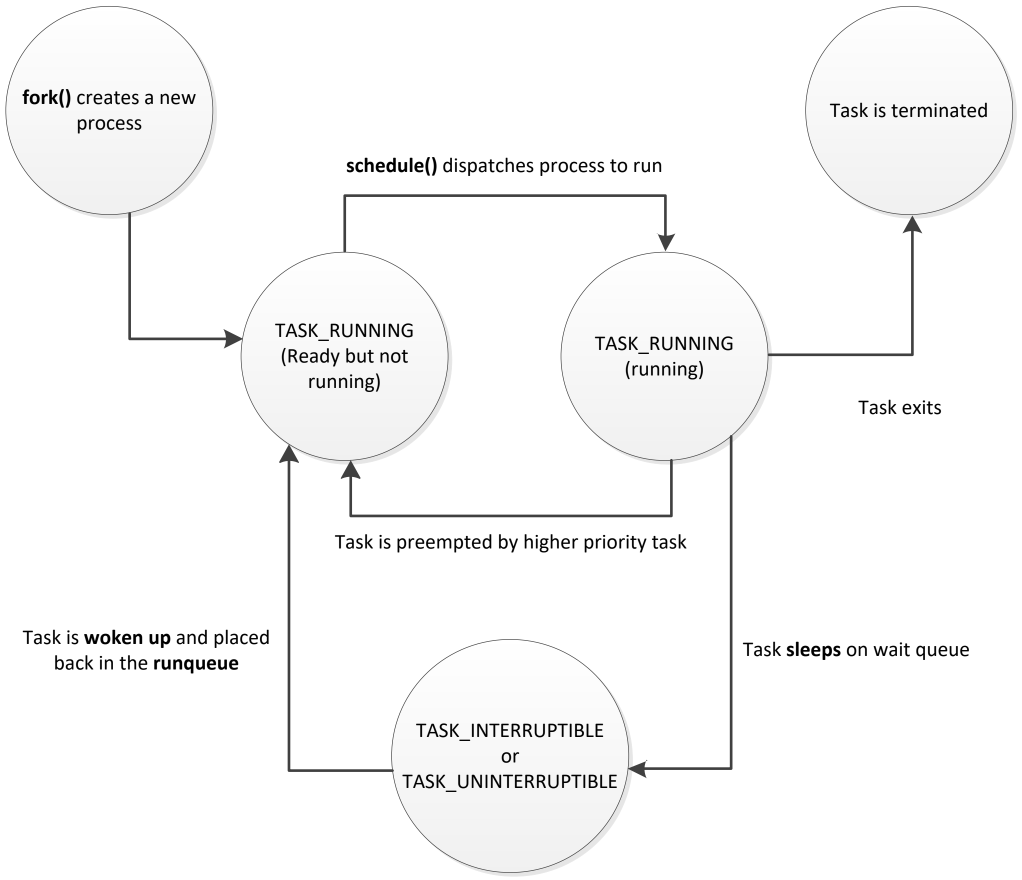
TASK_ZOMBIE: Process execution is terminated, but the parent process has not yet issued a **wait()** or **waitpid()** system call to return information about the dead process.[*] Before the wait()-like call is issued, the kernel cannot discard the data contained in the dead process descriptor because the parent might need it.

The value of the state field can be set using simple assignment, i.e,

```
p->state = TASK_RUNNING;
```

or using the macros **set_current_state** and **set_task_state**.

The life cycle of a process is shown in the following diagram:


fork() creates a new process
Task is terminated
schedule() dispatches process to run
TASK_RUNNING (Ready but not running)
TASK_RUNNING (running)
Task exits
Task is preempted by higher priority task
Task is **woken up** and placed back in the **runqueue**
Task **sleeps** on wait queue
TASK_INTERRUPTIBLE or TASK_UNINTERRUPTIBLE

Task Scheduling

Note: The scheduling algorithm, called the O(1) scheduler, in our kernel (version 2.4) is taken from a kernel version 2.6. If you want to look online for information on the algorithm, look for information relevant for kernel version 2.6 (or use the algorithm's name).

The scheduling algorithm is implemented in "kernel/sched.c", and most of the logic is in the **scheduler_tick** and **schedule** functions. Linux scheduling is based on "time sharing" technique. The CPU time is divided into *slices*, one for each runnable process (processes with the **TASK_RUNNING** state). A duration of the time slice depends on the priority of the process and ranges between 10ms to 300ms. Each CPU runs only one process at a time. The kernel keeps track of time using timer interrupts. When the time slice of the currently running process expires, the kernel scheduler is invoked and another task is set to run for the duration of its time slice. Switching between tasks is done through **context** switch. Switching of the currently running task can also occur before the expiration of its time slice. This can occur due to interrupts that wake up processes with higher priority or when the currently running process yields execution to the kernel (e.g. **blocks** or **sleeps**).

The kernel holds all runnable processes (processes with the **TASK_RUNNING** state) in a data structure called a **runqueue**. The runnable processes are further divided to two arrays: processes that have yet to exhaust their time slice are in the **active** array, processes those whose time slice has expired are in the **expired** array. Each array is a collection of linked lists, one for each possible priority. The **runqueue** also points to the current running process (which obviously resides in the **active** array). Each time the **schedule()** function is called it selects the next running process by taking the first process from the first non-empty list in the **active** array. Once the time slice of a process expires it is moved to the **expired** array. When the **active** is empty, the **expired** and **active** arrays are switched, and the **active** processes are assigned new time slices.

Kernel Timing

The kernel keeps track of time using the *timer interrupt*. The timer interrupt is issued by the system timer (implemented in hardware). The period of the system timer is called 'tick'. The *timer interrupt* advances the tick counter (called **jiffies**), and initiates time dependent activities in the kernel (decrease the time slice of the current running process, wake up processes that **sleep** waiting for a timer event etc.). The **jiffies** variable (defined in "include/linux/sched.h") counts the system 'tick's event since start up. The 'tick' period duration depends on the specific linux version. The **HZ** variable is use for converting the 'tick's to seconds:

$$\boldsymbol{time\ in\ seconds = \frac{jiffies}{HZ}}$$